STATISTICS IN MEDICINE
Statist. Med. **18**, 3021–3036 (1999)

AN EFFICIENT DESIGN FOR VERIFYING DISEASE OUTCOME STATUS IN LARGE COHORTS WITH RARE EXPOSURES AND LOW DISEASE RATES

WARREN B. BILKER[1*], JESSE A. BERLIN[1], MITCHELL H. GAIL[2], AND BRIAN L. STROM[1]

[1] *University of Pennsylvania, Department of Biostatistics and Epidemiology and Center for Clinical Epidemiology and Biostatistics, Blockley Hall, 423 Guardian Drive, Philadelphia, PA 19104-6021, U.S.A.*
[2] *National Cancer Institute, Bethesda, MD 20892, U.S.A.*

SUMMARY

Cohort studies require the use of large samples when the risk of the event is very low. Databases that are large and population-based, such as Medicaid files, are frequently used for cohort studies, since they provide access to the large samples required for adequate statistical power at a relatively affordable cost. Epidemiologic studies using these databases typically require verification of reported diagnoses, however, because of the potential for errors in disease reporting. When exposure prevalence is also low, as in many pharmacoepidemiologic investigations of drug toxicity, there are few exposed cases compared to the number of unexposed cases. Verification of all unexposed presumptive cases through medical records is costly. We investigate the statistical efficiency of a design in which all exposed cases but only a subsample of the unexposed cases are verified. We show that good efficiency can usually be achieved with a small subsample of unexposed cases. Published in 1999 by John Wiley & Sons, Ltd. This is a US Government work and is in the public domain in the United States.

1. INTRODUCTION

When the risk of an event is very low, cohort studies require the use of large samples. In this context, even case-control studies can require large samples when the prevalence of exposure is also low, as in the study of many pharmaceutical exposures. A convenient approach is to use large, population-based databases, such as Medicaid files, because they provide easy and relatively affordable access to the sample sizes required for adequate statistical power.[1] Although such databases pose methodologic problems, they are commonly used.

For example, pharmacoepidemiology studies are commonly performed after new drugs are marketed. Often, when these are conducted, the adverse drug reactions of interest are both serious and so rare that rates of such events cannot be determined precisely to permit reliable comparisons between marketed drugs based only on pre-marketing studies, which usually have 500–3000 subjects. Post-marketing studies are typically based on very large databases, commonly exceeding 100,000 subjects. For example, Strom *et al.*[2] studied oral contraceptives as a risk factor for gall-bladder disease using a retrospective cohort design. The database included the complete set of Medicaid

* Correspondence to: Warren B. Bilker, University of Pennsylvania, Department of Biostatistcs and Epidemiology and Center for Clinical Epidemiology and Biostatistics Blockley Hall, Room 601, 423 Guardian Drive, Philadelphia, PA 19104-6021, U.S.A. E-mail: wbilker@cceb.upenn.edu

Received July 1998
Accepted January 1999

Published in 1999 by John Wiley & Sons, Ltd.

billing data from the states of Michigan and Minnesota for the years 1980–1981 and permitted a comparison of 139,943 oral contraceptive users with 341,478 non-users. The nested case-control study is another common design used in pharmacoepidemiology studies.[3]

One of the primary methodologic concerns with population-based database studies is the potential for errors in the diagnoses reported in the database. Thus, the performance of pharmacoepidemiologic studies using these large databases usually requires verification of diagnoses reported in the database,[1] which may be accomplished, for example, by obtaining medical records of the presumptive cases identified in the database. Typically, the number of individuals exposed to the drug of interest who develop disease is very small because both the prevalence of exposure and the incidence of disease are low, whereas the number of unexposed individuals who develop disease may be much larger, because of the large pool of unexposed people. Retrieving medical records for all of the unexposed individuals with disease can be costly. At a cost of perhaps $100–$200 per record, retrieval of medical records is often a major expense and may limit the feasible study size. Therefore, we investigated whether verifying the disease status of all exposed cases but only a random subsample of unexposed cases might result in substantial cost savings with only a small loss of statistical efficiency relative to complete verification.

Brenner and Gefeller[4] examined a similar design for estimating relative risks, but they used the same sampling fraction for verifying the disease status of exposed and unexposed presumptive cases. Our emphasis differs in that we allow different sampling fractions for verifying the disease status of exposed and unexposed presumptive cases. This is especially useful in the common setting for pharmacoepidemiology studies, where most cases are unexposed. The proposed method yields smaller variances for relative risk estimates than the method of Brenner and Gefeller for typical scenarios in pharmacoepidemiologic studies. Moreover, the methods we present can be used to study risk differences and other measures of exposure effect, as well as relative risks.

We assume that all cases of the disease are detected (100 per cent sensitivity) but that some individuals may be misclassified as diseased because of imperfect specificity of the diagnoses. Thus, the methods we propose are appropriate for the study of serious conditions, such as bleeding peptic ulcers, that require hospitalization and have a high probability of detection. In order to adapt our methods for less serious diseases that might not always be detected (sensitivity less than 100 per cent), we would need to validate not only cases but also non-cases. If such a condition were common, its incidence would be estimated more accurately from pre-marketing studies. If such a condition were rare, a post-marketing study would probably not be warranted. Therefore, we focus on studies in which events are both serious and rare, and for which the assumption of 100 per cent sensitivity is reasonable.

Following a presentation of notation and methods, we give results on the asymptotic relative efficiency comparing full verification of all presumptive cases to verification of a subsample of presumptive cases with different sampling fractions for the unexposed and exposed presumptive cases. We then present the asymptotic relative efficiency of the proposed method compared to the method of Brenner and Gefeller. Simulations are presented that confirm that the confidence intervals for relative risks estimated according to the proposed procedures have nominal coverage. Procedures are illustrated by numerical examples, including a realistic example based on a study to determine the risk of liver disease following treatment with non-steroidal anti-inflammatory drugs.

Table I. Notation

		True cases	Falsely diagnosed cases	Total presumptive cases	Size of subsample from presumptive cases	Number of true cases in the subsample
Unexposed	N_0	D_0	W_0	$T_0 = D_0 + W_0$	$M_0 \leqslant T_0$	X_0
Exposed	N_1	D_1	W_1	$T_1 = D_1 + W_1$	$M_1 \leqslant T_1$	X_1

Quantities of interest: $p_0 = D_0/N_0$; $p_1 = D_1/N_1$

2. NOTATION AND METHODS

Let N_0 and N_1 be the numbers of unexposed and exposed members of the population, respectively, and let other quantities be defined as in Table I. We are interested in $p_0 = P(\text{true disease}|\text{unexposed})$ and $p_1 = P(\text{true disease}|\text{exposed})$. Usual risk measures such as the risk difference $p_1 - p_0$ or relative risk p_1/p_0 can be based on estimates of p_0 and p_1.

A two-stage approach to estimating p_0 is based on $P(\text{true disease or false disease}|\text{unexposed}) \times P(\text{true disease}|(\text{true disease or false disease}) \text{ and unexposed})$. This equation holds because the subset of patients with 'true disease' is included in the patients with 'true disease or false disease', provided all individuals with true disease are identified in the database as having the disease of interest (100 per cent sensitivity). It is also assumed that the exposure is measured without error. Letting $p_0^\star = P(\text{true disease or false disease} \mid \text{unexposed})$ and $\pi_0 = P(\text{true disease} \mid (\text{true disease or false disease}) \text{ and unexposed})$, we obtain $p_0 = p_0^\star\ \pi_0$. The corresponding estimate is $\hat{p}_0 = \hat{p}_0^\star\ \hat{\pi}_0$, where $\hat{p}_0^\star = (D_0 + W_0)/N_0$, $\hat{\pi}_0 = X_0/M_0$, and, as defined in Table I, D_0 is the number of true cases who are unexposed, W_0 is the number of falsely diagnosed cases who are unexposed, and X_0 is the number of true cases in the subsample of M_0 unexposed cases selected for verification. An exactly analogous procedure is used to obtain $\hat{p}_1$. The quantities π_0 and π_1 are referred to as the 'positive predictive values'.

An estimate of the variance of $\hat{p}_0$, $\text{var}(\hat{p}_0) = \text{var}(\hat{p}_0^\star\ \hat{\pi}_0)$ can be obtained using the multivariate form of the 'delta method' and the fact that $\hat{p}_0^\star$ and $\hat{\pi}_0$ are uncorrelated, which is proved in the Appendix. Define the proportion of unexposed presumptive cases that are selected in the sample to be verified as f_0, where $f_0 = M_0/T_0$. Then, the variance of $\hat{p}_0$ is

$$\begin{aligned} \text{var}(\hat{p}_0) &= \pi_0^2 \text{var}(\hat{p}_0^\star) + (p_0^\star)^2 \text{var}(\hat{\pi}_0) \\ &= \frac{\pi_0^2 p_0^\star\ (1 - p_0^\star)}{N_0} + \frac{p_0^\star\ \pi_0(1 - \pi_0)}{f_0 N_0}. \end{aligned} \tag{1}$$

An estimate of $\text{var}(\hat{p}_0)$ is obtained by substituting $\hat{p}_0^\star$ and $\hat{\pi}_0$ for $p_0^\star$ and π_0. The derivation of this variance estimator is provided in the Appendix. An analogous result can be obtained for $\text{var}(\hat{p}_1)$, where $f_1 = M_1/T_1$ is the fraction of exposed presumptive cases that are selected in the sample to be verified. The intent is to verify the disease status of all exposed cases, which leads to $f_1 = 1$. In practice, however, it is rare that all medical records requested for verification are received, and $f_1 < 1$ is therefore common. The validity of our results depends on the assumption that the availability of medical records among the T_0 unexposed presumptive cases is unrelated to actual (true) disease status, and, in particular, we assume that the M_0 unexposed subjects whose diagnoses are reviewed are representative of all T_0 unexposed presumptive cases. Similar assumptions are made for T_1 and M_1.

3. RESULTS

3.1. Asymptotic relative efficiency of method for log($\widehat{RR}$) relative to full verification

Suppose we want to compute the variance of the estimated log relative risk, $\log(\hat{p}_1/\hat{p}_0)$, adjusted for misclassification. By the delta method,

$$\text{var}(\log(\hat{p}_1/\hat{p}_0)) = \frac{1}{p_0^2}\text{var}(\hat{p}_0) + \frac{1}{p_1^2}\text{var}(\hat{p}_1) \tag{2}$$

where the cross-product term is 0 since $\hat{p}_0$ and $\hat{p}_1$ are uncorrelated. A similar approach could be applied to the odds ratio, risk difference, or any other regular function of p_0 and p_1. Estimates are obtained by substituting $\hat{p}_0$ and $\hat{p}_1$ for p_0 and p_1.

We define the asymptotic relative efficiency (ARE) from verifying $M_0 = f_0 T_0$ unexposed cases and $M_1 = f_1 T_1$ exposed cases as the ratio of the variance of the log relative risk with complete verification, $f_0 = 1$ and $f_1 = 1$, to the variance with fractional verification, $f_0 \leqslant 1$ and $f_1 \leqslant 1$. The ARE is obtained by substituting the results for $\text{var}(\hat{p}_0)$ and $\text{var}(\hat{p}_1)$ (equation (1)) into equation (2). It follows that the ARE is:

$$\frac{\frac{1}{p_0^2}\left[\frac{\pi_0^2 p_0^\star(1-p_0^\star)}{N_0} + \frac{p_0^\star \pi_0(1-\pi_0)}{N_0}\right] + \frac{1}{p_1^2}\left[\frac{\pi_1^2 p_1^\star(1-p_1^\star)}{N_1} + \frac{p_1^\star \pi_1(1-\pi_1)}{N_1}\right]}{\frac{1}{p_0^2}\left[\frac{\pi_0^2 p_0^\star(1-p_0^\star)}{N_0} + \frac{p_0^\star \pi_0(1-\pi_0)}{f_0 N_0}\right] + \frac{1}{p_1^2}\left[\frac{\pi_1^2 p_1^\star(1-p_1^\star)}{N_1} + \frac{p_1^\star \pi_1(1-\pi_1)}{f_1 N_1}\right]}.$$

Define $k = M_0/M_1$, the ratio of the number of presumptive cases in the unexposed subsample relative to the number in the exposed subsample. If $p_0^\star$ is small, then $p_0^\star(1-p_0^\star) \doteq p_0^\star$. Similarly, if $p_1^\star$ is small, then $p_1^\star(1-p_1^\star) \doteq p_1^\star$. Under these assumptions, the above representation for ARE simplifies to

$$\text{ARE} \doteq 1 - \frac{\pi_0 k + \pi_1 - \pi_0\pi_1((1-f_0) + k(1-f_1)) - \pi_1 f_0 - \pi_0 f_1 k}{\pi_0 k + \pi_1 - \pi_0\pi_1((1-f_0) + k(1-f_1))} \tag{3}$$

The ARE is 1 with full verification, $f_0 = f_1 = 1$, and falls below 1 when sampling the unexposed cases. We use equation (3) in the following discussion of ARE.

An important question in applying this methodology is how to achieve a particular ARE. The ARE depends on the sampling fractions, f_0 and f_1, the positive predictive values, π_0 and π_1, and the sampling ratio, k. Contour plots are presented showing the combinations of f_0, $\pi_0 = \pi_1$ and k that yield a specified ARE, with a specified f_1. Here, the positive predictive values for exposed and unexposed presumptive cases are taken to be equal, $\pi_0 = \pi_1$. Figure 1 is a contour plot showing the relationship between f_0, $\pi_0 = \pi_1$ and k for ARE = 0·8 and complete sampling of exposed presumptive cases, $f_1 = 1$. The value of k is denoted at the base of each contour line. For example, to achieve an ARE of 80 per cent when verifying equal numbers of exposed and unexposed cases ($k = 1$), one only needs to sample 24 per cent of the unexposed presumptive cases ($f_0 = 0{\cdot}24$) when the positive predictive values, $\pi_0 = \pi_1$, are 0·60. The ARE will exceed 80 per cent if either f_0, $\pi_0 = \pi_1$, or k are increased. Thus, for any fixed k, the ARE exceeds 80 percent if one moves outward (northeast) from the contour plot in Figure 1. Likewise, increasing k allows one to achieve an ARE of 80 per cent with smaller values of $\pi_0 = \pi_1$ or f_0. Thus, an ARE of 80 per cent is achieved with $f_0 = 0{\cdot}04$ and $\pi_0 = \pi_1 = 0{\cdot}60$ when $k = 1{\cdot}5$, and with $f_0 = 0{\cdot}13$ and $\pi_0 = \pi_1 = 0{\cdot}40$ when $k = 2$. Larger ratios of sampled unexposed to exposed cases (k) are needed for smaller values of positive predictive value, $\pi_0 = \pi_1$.

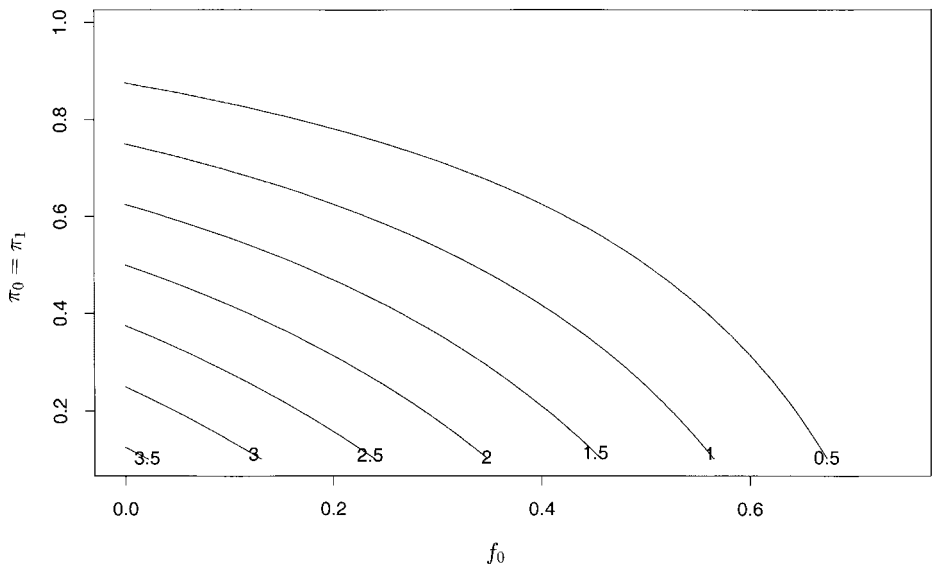

Figure 1. Contour plot for ARE of 80 per cent with $f_1 = 1{\cdot}0$. Each contour represents a different value of $k = M_0/M_1$

It is not always possible to validate all of the exposed presumptive cases (for example, missing records, access denial). The ARE above compares the variance based on complete verification of all exposed presumptive cases and a subsample of all unexposed presumptive cases to the variance obtained from complete verification. In cases where complete verification of all exposed presumptive cases is not possible, it is more meaningful to consider the ARE comparing the variance based on verification of the maximum achievable fraction of verified exposed presumptive cases and a subsample of all unexposed presumptive cases to the variance based on verification of the maximum achievable fraction of verified exposed presumptive cases and complete verification of all unexposed presumptive cases. This quantity is referred to as the 'adjusted ARE'. The 'adjusted ARE' compares the variance with a given maximum achievable $f_1 < 1$, $f_1(\max)$, and $f_0 = 1$ to the variance for the same $f_1(\max)$ and for $f_0 < 1$. This quantity can be obtained as the ratio of the ARE based on the desired $f_0 < 1$ with $f_1 = f_1(\max)$ to the ARE based on the desired $f_0 = 1$ with $f_1 = f_1(\max)$. In the example contour plots in Figures 1–3, the contour plots of 'adjusted ARE', when $f_1 < 1$, are very similar to those shown in Figures 1–3, indicating that for a fixed number of verified exposed cases, one can achieve good 'adjusted ARE' without sampling many more unexposed than exposed cases.

Similar results are found with a contour plot corresponding to an ARE of 90 per cent and $f_1 = 1$ (Figure 2). With the positive predictive value of 0·60 and $k = 1$, a sampling fraction, $f_0 = 0{\cdot}57$ would be required to achieve a 90 per cent ARE. If $k = 3$, then a sampling fraction, $f_0 = 0{\cdot}13$ is sufficient. In general, as k increases a smaller sampling fraction, f_0, is required to achieve $\text{ARE} = 0{\cdot}9$ for fixed $\pi_0 = \pi_1$. Also, as k increases, an $\text{ARE} = 0{\cdot}9$ can be achieved with decreasing values of $\pi_0 = \pi_1$ for fixed f_0.

As discussed above, it is not always possible to validate all of the exposed presumptive cases. A contour plot with $f_1 = 0{\cdot}9$ and an ARE of 80 per cent is displayed in Figure 3. Comparison of Figure 3 with Figure 1 shows that there is a trade-off between f_0, f_1 and k. A reduction in f_1

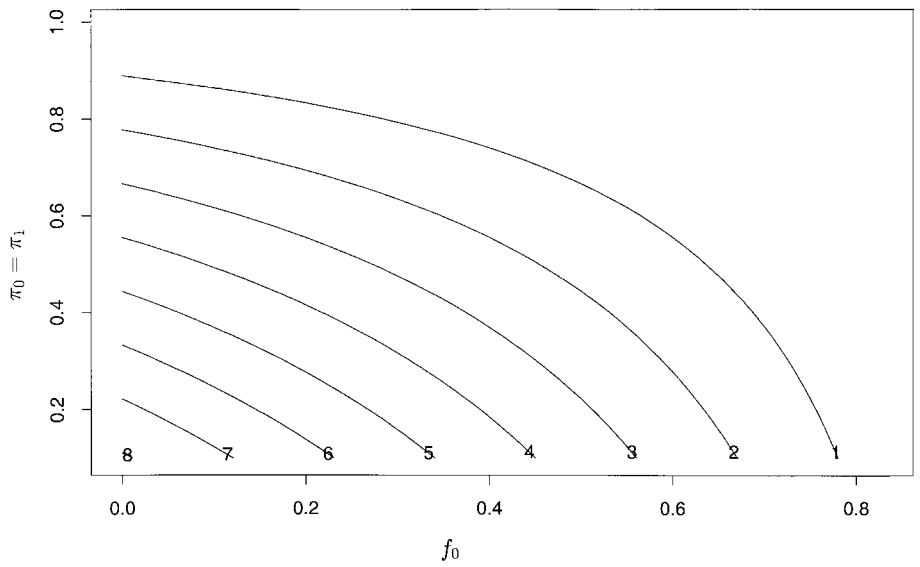

Figure 2. Contour plot for ARE of 90 per cent with $f_1 = 1{\cdot}0$. Each contour represents a different value of $k = M_0/M_1$

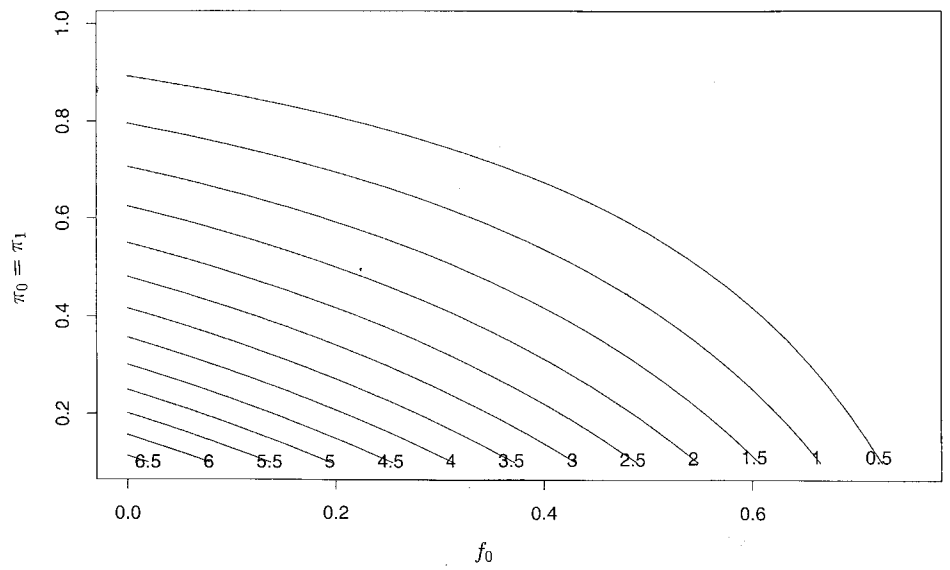

Figure 3. Contour plot for ARE of 80 per cent with $f_1 = 0{\cdot}9$. Each contour represents a different value of $k = M_0/M_1$

requires an increase in f_0 and/or k to maintain ARE $= 0{\cdot}8$. For example, with $f_1 = 1{\cdot}0$, an ARE of 80 per cent requires $f_0 = 0{\cdot}12$ when $\pi_0 = \pi_1 = 0{\cdot}40$ and $k = 2$ (Figure 1). However, if $f_1 = 0{\cdot}9$, an ARE of 80 per cent requires $f_0 = 0{\cdot}32$ when $\pi_0 = \pi_1 = 0{\cdot}40$ and $k = 2$ (Figure 3).

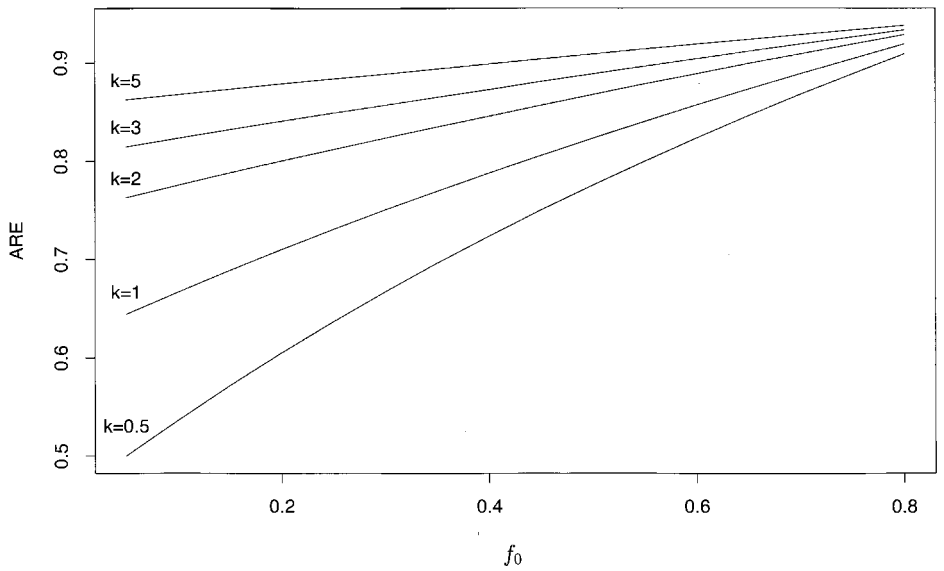

Figure 4. Plot of ARE against f_0 for different values of $k = M_0/M_1$, with $\pi_0 = \pi_1 = 0{\cdot}5$ and $f_1 = 1$

The ARE is plotted against f_0 for various values of k with $f_1 = 1$ and with $\pi_0 = \pi_1 = 0{\cdot}5$ in Figure 4. Figure 5 shows similar plots for $\pi_0 = \pi_1 = 0{\cdot}8$. It is seen that the ARE increases with increasing f_0. As k increases, the increase in ARE that results from a given increase in f_0 diminishes. The ARE also increases with increasing $\pi_0 = \pi_1$ for all fixed values of f_0 and k.

The important message is that for large population-based databases, 80 per cent efficiency or higher can be achieved by verifying all or a large proportion of exposed presumptive cases and a relatively small fraction of all unexposed presumptive cases. As the fraction of exposed presumptive cases that are verifiable decreases, the fraction of unexposed presumptive cases required and/or the ratio of the number of presumptive cases in the unexposed subsample to the number of presumptive cases in the exposed subsample need to be increased to achieve a specified ARE. For a fixed limited number of available exposed cases, however, one can achieve good 'adjusted ARE' by sampling modest numbers of unexposed cases.

3.2. Asymptotic relative efficiency of method for $\log(\widehat{RR})$ relative to the approach of Brenner and Gefeller

It is seen that good ARE can be obtained with verification of a small subsample of the unexposed presumptive cases. The method of Brenner and Gefeller (BG) also allows for verification of a subsample. The ARE for the BG method is a special case of the ARE formula presented here (equation (3)), with $f_0 = f_1 < 1$. In this section, we compare the efficiency of the method proposed here (BBGS) with the method of BG. We fix the total number of presumptive cases to be verified to be the same in both methods. The total number of cases verified using the BBGS method is $M_0 + M_1 = f_0 T_0 + f_1 T_1$. Defining $f = (M_0 + M_1)/(T_0 + T_1) = (f_0 T_0 + f_1 T_1)/(T_0 + T_1)$, we note that the BG method will sample the same number of cases as the BBGS method provided the BG method samples the same proportion, f, of presumptive exposed and unexposed cases. The

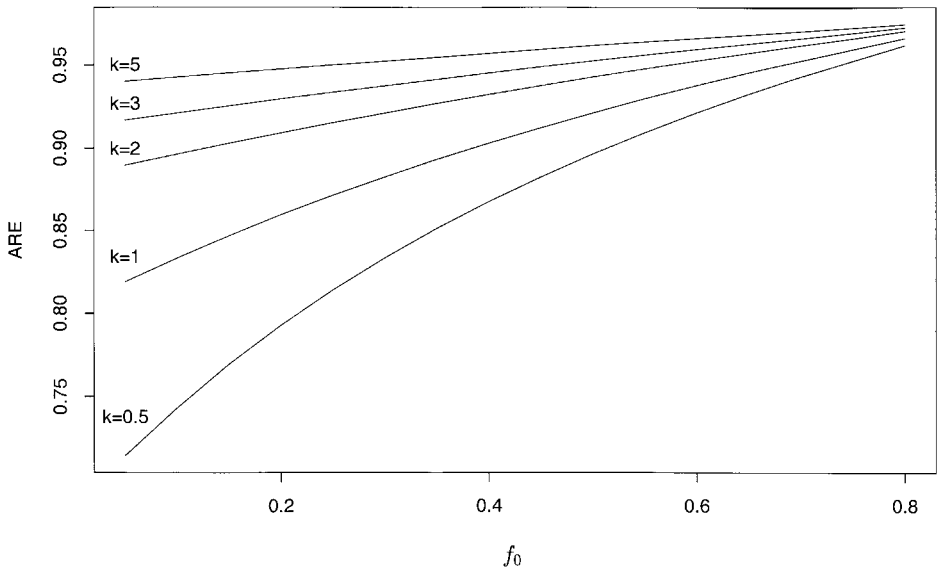

Figure 5. Plot of ARE against f_0 for different values of $k = M_0/M_1$, with $\pi_0 = \pi_1 = 0{\cdot}8$ and $f_1 = 1$

pertinent question is then, what is the relative performance of the two methods when the same total number of presumptive cases are verified, and thus the cost of verification is the same. By definition, $k = f_0 T_0 / f_1 T_1$ which yields $T_0 = (f_1 T_1 k)/f_0$. Substituting this result into f yields $f = f_0 f_1 (k+1)/(f_1 k + f_0)$. The ARE comparing full verification to the BG method is obtained by substituting $f_0 = f_1 = f$ into (3). The relative efficiency, RE, comparing BG to BBGS is obtained by taking the ratio of the variance of the BBGS method with sampling fractions $f_0 < 1$ and $f_1 = 1$ to the variance of the BG method with the common sampling fraction, $f = f_0(f_1 k + 1)/(f_1 k + f_0)$. After some algebraic simplification and use of the assumptions that $p_0^{\star}$ and $p_1^{\star}$ are both small, that were used in deriving (3), RE can be shown to be

$$\mathrm{RE} \doteq \frac{f_0(1+k)(k\pi_0 + \pi_1 - \pi_0\pi_1 + f_0\pi_0\pi_1)}{(f_0+k)(k\pi_0 + f_0\pi_1 - k\pi_0\pi_1 + f_0 k\pi_0\pi_1)}. \tag{4}$$

Here, RE less than 1 indicates that BBGS has smaller variance than BG for the same number of presumptive cases verified, and thus BG is less efficient than BBGS.

Since the positive predictive values (PPV) for the exposed and unexposed groups are commonly the same or similar, we first consider the case where $\pi_0 = \pi_1$. In this case, it can be shown that $\mathrm{RE} \leqslant 1$ when $k \geqslant f_0^{1/2}$ or equivalently when $f_0 > (T_1/T_0)^2$. Since the BBGS method is proposed here for scenarios where there are many more unexposed presumptive cases than exposed presumptive cases, $(T_1/T_0)^2$ will be small and thus this condition will usually be met. Thus, BBGS will usually be more efficient than BG. Typical examples of RE are shown in Table II.

The impact of unequal values of π_0 and π_1 depends on which group has the larger PPV (Table II). For any fixed π_0, as π_1 decreases relative to π_0, RE of BG decreases. Thus, the BG method becomes less efficient relative to the BBGS method as π_1 decreases for fixed values of π_0. If π_1 is large and π_0 is much smaller, BG can be more efficient than BBGS, as indicated by the entry with $\pi_0 = 0{\cdot}5$ and $\pi_1 = 0{\cdot}9$ in Table II. Such a scenario would occur if there were

Table II. Examples of ARE comparing BBGS and BG methods, RE

$T_0 = 500,\ T_1 = 25$ $\pi_0 = 0{\cdot}7,\ f_0 = 0{\cdot}3$ $k = 6$		$T_0 = 500,\ T_1 = 50$ $\pi_0 = 0{\cdot}6,\ f_0 = 0{\cdot}3$ $k = 3$		$T_0 = 500,\ T_1 = 100$ $\pi_0 = 0{\cdot}5,\ f_0 = 0{\cdot}3$ $k = 1{\cdot}5$	
π_1	RE	π_1	RE	π_1	RE
0·2	0·391	0·1	0·433	0·1	0·520
0·5	0·516	0·3	0·576	0·3	0·703
0·7	0·646	0·6	0·711	0·5	0·847
0·8	0·736	0·7	0·798	0·7	0·928
0·9	0·851	0·9	0·902	0·9	1·016

a suspected link between exposure and disease and physicians tended to base their diagnosis, in part, on the exposure status. We would not expect to see such a large difference between π_0 and π_1 in most applications, however.

3.3. Simulations to study coverage of the 95 per cent confidence interval on $\log(p_1/p_0)$

We conducted simulations to study the properties of asymptotic methods for realistic sample sizes and parameter values. To examine the coverage of the 95 per cent confidence interval on $\log(p_1/p_0)$, constructed as $\log(\hat{p}_1/\hat{p}_0) \pm 1{\cdot}96\{\widehat{\text{var}}(\log(\hat{p}_1/\hat{p}_0))\}^{1/2}$, we defined the relative risk $p_1/p_0 = \text{RR}$ and expressed π_0 and π_1 in terms of specificities s_0 and s_1, respectively, since these are typically more readily available than positive predictive values. To be precise, s_0 is the probability of no diagnosis of disease in unexposed non-diseased individuals and s_1 is defined similarly for exposed individuals. It follows that $\pi_0 = p_0/[p_0+(1-p_0)(1-s_0)]$ and $\pi_1 = p_1/[p_1+(1-p_1)(1-s_1)]$.

Four sets of simulations were performed, corresponding to various choices of RR, s_0, and s_1. In each set, N_1 was fixed at 10,000 exposed individuals and N_0 at 990,000 unexposed, while six different values of p_0 (0·001, 0·002, 0·005, 0·010, 0·020, 0·100) and 15 values of f_0 (0·001, 0·005, 0·01, 0·05, 0·10, 0·15, 0·2, 0·3, 0·4, 0·5, 0·6, 0·7, 0·8, 0·9, 1·0) were used. For each of the $4 \times 6 \times 15 = 360$ combinations studied, 10,000 independent data sets were generated. If the coverage is nominal at 95 per cent, the 95 per cent confidence interval for the coverage estimate would be $0{\cdot}950 \pm 0{\cdot}0043$.

To generate each sample, we recall that the N_0 unexposed cases fall into three categories: true positives (that is, truly diseased and presumed to have the disease); true negatives, and false positives, because by assumption there are no false negatives. A value for T_0, the number of unexposed presumptive cases, was randomly selected from a binomial distribution with parameters $p_0^{\star}$ and N_0. Conditional on T_0, the number of true positive cases, D_0, was randomly selected from a binomial distribution with index T_0 and probability π_0. The number of false positive cases was then computed as $W_0 = T_0 - D_0$. The next step was to determine the make-up of the subsample. The number of presumptive cases in the subsample is $M_0 = f_0 T_0$ (rounded up). The number of unexposed true positive cases selected in the subsample, X_0, was randomly selected from a hypergeometric distribution with indices (urn sizes) D_0 and W_0 and with a total of M_0 units selected out of the $T_0 = D_0 + W_0$ units. That is, X_0 units were taken to be true positives out of the M_0 unexposed presumptive cases selected in the subsample. An analogous procedure was followed for generating the random values for the exposed cases in the simulated data sets.

Table III. Example 1 information

		True cases	Falsely diagnosed cases	Total presumptive cases	Size of subsample from presumptive cases	Number of true cases in the subsample
Unexposed	$N_0 = 990,000$	$D_0 = 990$	$W_0 = 990$	$T_0 = 1980$	$M_0 = 150$	$X_0 = 75$
Exposed	$N_1 = 10,000$	$D_1 = 50$	$W_1 = 100$	$T_1 = 150$	$M_1 = 150$	$X_1 = 50$

Quantities of interest:
$p_0 = D_0/N_0 = 0{\cdot}001$
$\hat{p}_0^{\star} = (D_0 + W_0)/N_0 = (990 + 990)/990,000 = 0{\cdot}002$
$\hat{\pi}_0 = X_0/M_0 = 75/150 = 0{\cdot}5,\ \hat{p}_0 = \hat{p}_0^{\star} \times \hat{\pi}_0 = 0{\cdot}002 \times 0{\cdot}5 = 0{\cdot}001$
$p_1 = D_1/N_1 = 0{\cdot}005$
$\hat{p}_1^{\star} = (D_1 + W_1)/N_1 = (50 + 100)/10,000 = 0{\cdot}015$
$\hat{\pi}_1 = X_1/M_1 = 50/150 = 1/3,\ \hat{p}_1 = \hat{p}_1^{\star} \times \hat{\pi}_1 = 0{\cdot}015 \times (1/3) = 0{\cdot}005$
$k = M_0/M_1 = 150/150 = 1,\ f_0 = M_0/T_0 = 150/1980 = 0{\cdot}08$
Estimated relative risk $= \hat{p}_1/\hat{p}_0 = \hat{p}_1/\hat{p}_0 = 5/1 = 5$ (note that the uncorrected relative risk is $\hat{p}_1^{\star}/\hat{p}_0^{\star} = 15/2 = 7{\cdot}5$)

Table IV. Example 1: standard error of the log(RR)

M_0	$M_1 = 150 = T_1$					$M_1 = 90$			
	f_0	k	SE($\log \widehat{RR}$)	ARE	RE	k	SE($\log \widehat{RR}$)	ARE	Adjusted ARE*
150	0·08	1	0·165	0·773	0·258	1·67	0·190	0·583	0·829
300	0·15	2	0·154	0·881	0·328	3·33	0·181	0·642	0·913
600	0·30	4	0·149	0·948	0·477	6·67	0·176	0·676	0·962
1980	1·00	13·2	0·145	1·000	1·000	22	0·173	0·703	1·000

*Compared to $M_1 = 90,\ M_0 = 1980$

The observed coverages for the 95 per cent confidence intervals for RR = 5, $s_0 = 0{\cdot}998999$ and $s_1 = 0{\cdot}989950$ fell within the expected interval $0{\cdot}95 \pm 0{\cdot}0043 = (0{\cdot}9457, 0{\cdot}9543)$ in all but 5 out of the 90 cases examined (range: $0{\cdot}9443, 0{\cdot}9558$), compared to $0{\cdot}05 \times 90 = 4{\cdot}5$ expected. Thus, coverage was at near nominal levels despite the fact that some situations examined, such as $p_0 = 0{\cdot}001$ and $f_0 = 0{\cdot}001$, correspond to only 0·99 true unexposed cases expected in the sample and 50 true exposed cases. Very similar results were found for three other sets of simulations, (RR = 5, $s_0 = s_1 = 0{\cdot}95$), (RR = 2, $s_0 = s_1 = 0{\cdot}95$), (RR = 5, $s_0 = s_1 = 0{\cdot}80$), which had 4, 5 and 5 of 90 cases, respectively, falling outside the expected interval. We conclude that these procedures produce confidence intervals with near nominal coverage.

3.4. Numerical examples

Two numerical examples illustrate these calculations and the effect of using a non-exhaustive sample of $M_0 < T_0$ unexposed presumptive cases and a sample of $M_1 < T_1$ exposed presumptive cases. The first example is similar to those used in the simulations. The summary data for this example are presented in Table III. The exposed cases are sampled exhaustively ($M_1 = T_1$). The parameters $\hat{p}_0^{\star}$, $\hat{\pi}_0$, $\hat{p}_1^{\star}$, $\hat{\pi}_1$, $\hat{p}_0$ and $\hat{p}_1$ are all computed from the observable data.

Table IV presents estimates of the standard error of the log relative risk, $\log(\hat{p}_1/\hat{p}_0)$, for M_0 equal to 150, 300, 600 and 1980 (T_0) and for M_1 equal to 150 (T_1) or 90, with X_0 set to its expected

Table V. Example 2 information

		True cases	Falsely diagnosed cases	Total presumptive cases	Size of subsample from presumptive cases	Number of true cases in the subsample
Unexposed	$N_0 = 1,1000,000$	$D_0 = 181$	$W_0 = 519$	$T_0 = 700$	$M_0 = 379$ (actual)	$X_0 = 98$
Exposed	$N_1 = 71,000$	$D_1 = 20$	$W_1 = 58$	$T_1 = 78$	$M_1 = 35$ (actual)	$X_1 = 9$

Quantities of interest:
$p_0 = 0{\cdot}0001645$ (8-year cumulative incidence)
$\hat{p}_0^{\star} = (181 + 519)/1,100,000 = 0{\cdot}0006364$
$\hat{\pi}_0 = X_0/M_0 = 98/379 = 0{\cdot}259,\ \ \hat{p}_0 = 0{\cdot}0006364 \times 0{\cdot}259 = 0{\cdot}0001648$
$p_1 = 0{\cdot}0002817$ (8-year cumulative incidence)
$\hat{p}_1^{\star} = (20 + 58)/71,000 = 0{\cdot}0010986$
$\hat{\pi}_1 = 9/35 = 0{\cdot}257,\ \ \hat{p}_1 = 0{\cdot}0010986 \times 0{\cdot}257 = 0{\cdot}0002823$
$k = M_0/M_1 = 379/35 = 10{\cdot}83,\ \ F_0 = M_0/T_0 = 379/700 = 0{\cdot}51$
Estimated relative risk $= \hat{p}_1/\hat{p}_0 = 2{\cdot}8234/1{\cdot}6483 = 1{\cdot}7$ (note that the uncorrected relative risk is $\hat{p}_1^{\star}/\hat{p}_0^{\star} = 10{\cdot}986/6{\cdot}364 = 1{\cdot}7$, which is unbiased because the proportions of exposed and unexposed presumptive cases who are true cases are nearly identical)

value, $E(X_0) = 0{\cdot}5M_0$. The estimated ARE for each set of values relative to complete verification is also presented. Note that the AREs computed as ratios of the squared standard errors equal those calculated from equation (3). For $M_1 = T_1 = 150$, this example shows that $M_0 = 600$ ($f_0 = 0{\cdot}30$) yields an ARE of 0·948. The relative efficiency, RE, comparing BG to BBGS (equation (4)) is also presented. In all cases with less than full verification, BBGS is more efficient than BG. For $M_0 = M_1 = 150$, the RE is 0·258, indicating that the variance of the $\log(p_1/p_0)$ obtained using BBGS would be approximately one-fourth that from using BG. If only $M_1 = 90$ exposed cases can be verified, $M_0 = 600$ ($f_0 = 0{\cdot}30$) yields 68 per cent efficiency compared to an exhaustive sample with $f_0 = 1$ and $f_1 = 1$. However, a more relevant comparison is between the cases with $M_0 = 1980$ and $M_0 = 600$, both with $M_1 = 90$, because 90 is the largest possible number of exposed cases that can be verified. The column denoted 'Adjusted ARE' uses $M_1 = 90$ and $M_0 = 1980$ as the reference. The 'Adjusted ARE' for $M_0 = 600$ ($f_0 = 0{\cdot}30$) is 96 per cent.

The next example is based on a study of the risk of liver disease from non-steroidal anti-inflammatory drugs.[5,6] Although the study was originally performed as a case-control study using COMPASS, a large Medicaid database, we were able to obtain the data to perform the cohort analysis described below. All presumptive cases hospitalized with liver disease found over an 8-year period were identified for potential entry into the study. Medical records for all cases were requested to verify their diagnoses. In addition, the records were used to exclude cases whose liver disease was likely to be related to alcohol consumption or other underlying conditions. In this example, the investigators were able to obtain only 35 (45 per cent) of the 78 requested records for the exposed cases. Note also that the investigators observed $\hat{\pi}_0$ and $\hat{\pi}_1$, in this study, and we have set $D_0 = \hat{p}_0 T_0$, $D_1 = \hat{p}_1 T_1$, $W_0 = T_0 - D_0$ and $W_1 = T_1 - D_1$. The summary data for this example are presented in Table V. The parameters $\hat{p}_0^{\star}$, $\hat{\pi}_0$, $\hat{p}_1^{\star}$, $\hat{\pi}_1$, $\hat{p}_0$ and $\hat{p}_1$ are all computed based on the observable data.

Table VI provides estimates of the ARE and the standard error of $\log(\hat{p}_1/\hat{p}_0)$ for various values of M_0, with X_0 set to its expected value of $0{\cdot}259M_0$ and M_1 set to both the observed value of 35 and the complete sampling value of 78. First, consider the case with $M_1 = 78$. With $M_0 = 350$ ($f_0 = 0{\cdot}5$), an ARE of 93 per cent is attained. For $M_0 = 200$ ($f_0 = 0{\cdot}29$) the ARE is

Table VI. Example 2: standard error of the log(RR)

M_0	$M_1 = 78 = T_1$					$M_1 = 35$			
	f_0	k	SE(log $\widehat{\text{RR}}$)	ARE	RE	k	SE(log $\widehat{\text{RR}}$)	ARE	Adjusted ARE*
35	0·05	0·45	0·365	0·416	0·447	1	0·423	0·310	0·565
70	0·10	0·90	0·304	0·601	0·400	2	0·371	0·402	0·733
105	0·15	1·35	0·280	0·705	0·415	3	0·352	0·446	0·813
200	0·29	2·56	0·256	0·844	0·507	5·71	0·333	0·498	0·908
350	0·50	4·49	0·244	0·931	0·668	10	0·324	0·528	0·961
700	1	8·97	0·235	1·000	1·000	20	0·318	0·549	1·000

* Compared to $M_1 = 35$, $M_0 = 700$

Table VII. Example 1: standard error of the risk difference ($\hat{p}_1 - \hat{p}_0$)

M_0	$M_1 = 150 = T_1$		$M_1 = 90$		
	SE($\widehat{\text{Risk Difference}}$)	ARE	SE($\widehat{\text{Risk Difference}}$)	ARE	Adjusted ARE*
150	0·000710	0·988	0·000853	0·686	0·992
300	0·000708	0·994	0·000851	0·689	0·996
600	0·000707	0·998	0·000850	0·691	0·998
1980	0·000706	1·000	0·000849	0·692	1·000

* Compared to $M_1 = 90$, $M_0 = 1980$

84 per cent. The relative efficiency, RE, is also presented. For this example, the BBGS method is more efficient than the BG method in all cases with less than full verification. For $M_0 = 35$ and $M_1 = 78$, the RE is 0·447, indicating that the variance of the $\log(p_1/p_0)$ obtained using BBGS would be approximately one-half that from using BG. In this example, it was only possible to obtain medical records for verification for 35 exposed cases. Therefore, we computed the adjusted ARE relative to $M_1 = 35$ and $M_0 = 700$. The adjusted ARE for $M_0 = 200$ ($f_0 = 0{\cdot}29$) is 91 per cent, and for $M_0 = 105$ ($f_0 = 0{\cdot}15$) the adjusted ARE is 81 per cent.

In Tables VII and VIII, we present estimates of the standard error of the risk difference and AREs for the previous examples. The variance of the risk difference is estimated by the sum of the variances of $\hat{p}_0$ and $\hat{p}_1$. The patterns of results are similar to those for the log relative risk, except that the AREs for the risk differences are all consistently higher than for the log relative risk. In the first example (Table VII), $M_0 = 150$ ($f_0 = 0{\cdot}075$) and $M_1 = 150$ ($f_1 = 1{\cdot}0$) yield an ARE of 99 per cent. If M_1 is reduced to $M_1 = 90$ ($f_1 = 0{\cdot}6$) the ARE drops to 69 per cent; the adjusted ARE remains at 99 per cent, however. For the log relative risk, the case with $M_0 = 150$ and $M_1 = 150$ yields an ARE of 77 per cent (from Table IV). The case with $M_0 = 150$ and $M_1 = 90$ yields an ARE of 58 per cent, with an adjusted ARE of 83 per cent. Similar patterns of efficiencies are found for the example concerning liver disease (Table VIII). Thus, smaller sampling fractions would be required for the risk difference than for the log relative risk to achieve similar efficiencies.

Table VIII. Example 2: standard error of the risk difference ($\hat{p}_1 - \hat{p}_0$)

M_0	$M_1 = 78 = T_1$		$M_1 = 35$		
	SE($\widehat{\text{Risk Difference}}$)	ARE	SE($\widehat{\text{Risk Difference}}$)	ARE	Adjusted ARE ★
35	0·0000790	0·662	0·0000993	0·418	0·786
70	0·0000716	0·805	0·0000936	0·471	0·886
105	0·0000690	0·868	0·0000916	0·492	0·925
200	0·0000664	0·937	0·0000896	0·514	0·965
350	0·0000651	0·974	0·0000887	0·524	0·986
700	0·0000642	1·000	0·0000881	0·532	1·000

★ Compared to $M_1 = 35$, $M_0 = 700$

4. DISCUSSION

For cohort studies of uncommon exposures and rare disease outcomes, we have proposed verifying all exposed presumptive cases, or as many as can be verified, but only a fraction of unexposed presumptive cases. Our results generalize those of Brenner and Gefeller[4] who studied relative efficiencies, assuming that the same sampling fraction was used for both exposed and unexposed presumptive cases and confined attention to relative risks. Our methods also allow one to calculate efficiencies for risk differences and other measures of exposure effect. Brenner and Gefeller[4] showed that relative efficiency keeps increasing as the common sampling fraction increases, whereas we find sharply diminishing returns from increasing the sampling fraction for unexposed presumptive cases. For our subsampling approach, it is shown that in many situations, quite low sampling fractions for verification of the unexposed cases, in the 20–30 per cent range, can achieve high efficiency relative to exhaustive sampling for estimating the relative risk. Smaller sampling fractions would be required for the risk difference. This finding is similar in principle to the result that increasing the number of controls per case in a case-control study with a fixed number of cases yields diminishing increases in efficiency as the control to case ratio increases,[7] although the problems and mathematics differ. It is worth noting that our approach uses an internal validation sample and does not depend on external samples.

Our approach can lead to substantial reductions in the number of presumptive cases for whom the disease status is verified, resulting in substantial cost savings. Considering the same number of verified presumptive cases, our methods are also more efficient than using the common sampling fraction approach of Brenner and Gefeller for cohort studies of rare exposures and disease outcomes. Thus, equal efficiency can be achieved with a smaller number of cases verified using our methods, leading to substantial cost savings.

In using these methods, one should bear in mind the assumptions: (i) the database has 100 per cent sensitivity for detecting true cases, though specificity may be imperfect; (ii) the subsamples of unexposed presumptive cases and exposed presumptive cases are representative of all unexposed presumptive cases and all exposed presumptive cases, respectively, with respect to their probability of having true disease; and (iii) the exposure is measured without error.

As described in the introduction, assumption (i) is reasonable for the proposed applications of this method. Consider the case, however, where the sensitivities for detecting exposed and unexposed cases are less than perfect. If the sensitivity of the database for detecting unexposed cases is $g_0 < 1$ and that for detecting exposed cases is $g_1 < 1$, then the estimates of $\hat{p}_0$ and $\hat{p}_1$

will estimate $g_0 p_0$ and $g_1 p_1$, respectively. If $g_0 = g_1$, estimates of the relative risk or log relative risk remain unbiased, though the precision of these estimates may be diminished because fewer presumptive cases will be available for subsampling. Estimates of risk differences will remain biased even if $g_0 = g_1$, however.

It was assumed that the exposure information is measured without error. As discussed by Carson *et al.*,[1,8] a Food and Drug Administration funded validation study of a Medicaid database showed that the drug exposure data are of 'extremely high quality'. If, however, the exposure information is measured with error, all of the methods discussed will yield potentially invalid results. Biases resulting from the simultaneous misclassification of the exposure and disease status are addressed in Brenner, *et al.*[9]

APPENDIX

From the Delta method we obtain

$$\mathrm{var}(\hat{p}_0) = \pi_0^2 \,\mathrm{var}\,(\hat{p}_0^{\star}) + (p_0^{\star})^2 \,\mathrm{var}\,(\hat{\pi}_0) + 2\pi_0 p_0^{\star} \mathrm{cov}(\hat{p}_0^{\star} \hat{\pi}_0).$$

It is now shown that $\hat{p}_0^{\star} = T_0/N_0$ and $\hat{\pi}_0 = X_0/M_0$ are uncorrelated. The conditional covariance $\mathrm{cov}(\hat{p}_0^{\star}, \hat{\pi}_0 | D_0, W_0) = 0$, because $T_0 = D_0 + W_0$. Therefore

$$\begin{aligned}
\mathrm{cov}(\hat{p}_0^{\star}, \hat{\pi}_0) &= E[0] + \mathrm{cov}[E(\hat{p}_0^{\star}|D_0, W_0), E(\hat{\pi}_0|D_0, W_0)] \\
&= \mathrm{cov}\left[\frac{T_0}{N_0}, \frac{D_0}{D_0 + W_0}\right] \\
&= E\left(\frac{D_0}{N_0}\right) - E\left(\frac{T_0}{N_0}\right) E\left(\frac{D_0}{D_0 + W_0}\right) \\
&= p_0 - p_0^{\star}\pi_0 = 0.
\end{aligned}$$

The same arguments show $\hat{p}_1^{\star}$ is uncorrelated with $\hat{\pi}_1$.

Thus, $\mathrm{var}(\hat{p}_0) = \pi_0^2 \mathrm{var}(p_0^{\star}) + (p_0^{\star})^2 \mathrm{var}(\pi_0)$. The variances, $\mathrm{var}(\hat{p}_0^{\star})$ and $\mathrm{var}(\hat{\pi}_0)$, are required. $\mathrm{var}(\hat{p}_0^{\star}) = p_0^{\star}(1 - p_0^{\star})/N_0$ because $\hat{p}_0^{\star} = T_0/N_0$ and T_0 is binomial with index N_0 and probability $p_0^{\star}$. $\mathrm{var}(\hat{\pi}_0)$ can be derived in the following manner. Conditional on T_0, the number of true unexposed cases, D_0, is binomial with index T_0 and probability π_0. The total number of true unexposed cases in the sample, out of the M_0 total unexposed presumptive cases in the sample, is denoted by X_0. Thus, $\mathrm{var}(\hat{\pi}_0)$ is the variance of X_0/M_0 where M_0 subjects are selected at random without replacement from T_0 subjects and $M_0 = f_0 T_0$. Conditional on the total number of presumptive unexposed cases, T_0, and the true total number of unexposed cases, D_0, we sample $M_0 = f_0 T_0$ presumptive unexposed cases without replacement. Thus, X_0 follows a hypergeometric distribution. Therefore

$$\mathrm{var}(\hat{\pi}_0) = \mathrm{var}\left(\frac{X_0}{M_0}\right) = E\left[\mathrm{var}\left(\frac{X_0}{M_0}\middle| D_0, T_0\right)\right] + \mathrm{var}\left(E\left[\frac{X_0}{M_0}\middle| D_0, T_0\right]\right)$$

where

$$E\left[\frac{X_0}{M_0}\middle| D_0, T_0\right] = E\left[\frac{X_0}{f_0 T_0}\right] = \frac{D_0 f_0 T_0}{T_0 f_0 T_0} = \frac{D_0}{T_0}$$

and

$$\operatorname{var}\left(\frac{X_0}{M_0}\middle| D_0, T_0\right) = \frac{1}{(f_0 T_0)^2}\frac{D_0(T_0 - D_0) f_0 T_0 (1 - f_0) T_0}{T_0^2(T_0 - 1)}$$
$$= \frac{(1 - f_0)}{f_0}\frac{D_0(T_0 - D_0)}{T_0^2(T_0 - 1)} \doteq \frac{(1 - f_0)}{f_0}\frac{D_0(T_0 - D_0)}{T_0^3}.$$

It follows that

$$E\left[\operatorname{var}\left(\frac{X_0}{M_0}\middle| D_0, T_0\right)\right] \doteq \frac{(1 - f_0)}{f_0} E\left[E\left[\frac{D_0(T_0 - D_0)}{T_0^3}\middle| T_0\right]\right]$$
$$= \frac{(1 - f_0)}{f_0} E\left[\frac{1}{T_0^3}\{T_0^2\pi_0 - (T_0\pi_0(1 - \pi_0) + (\pi_0 T_0)^2)\}\right]$$
$$= \frac{(1 - f_0)}{f_0} E\left[\frac{\pi_0}{T_0} - \frac{\pi_0(1 - \pi_0)}{T_0^2} - \frac{\pi_0^2}{T_0}\right]$$
$$\doteq \frac{(1 - f_0)}{f_0}\frac{\pi_0(1 - \pi_0)}{p_0^{\star} N_0}$$

and

$$\operatorname{var}\left(E\left[\frac{X_0}{M_0}\middle| D_0, T_0\right]\right) = \operatorname{var}\left(\frac{D_0}{T_0}\right) = E\left[\operatorname{var}\left(\frac{D_0}{T_0}\middle| T_0\right)\right] + \operatorname{var}\left(E\left[\frac{D_0}{T_0}\middle| T_0\right]\right)$$
$$= E\left[\frac{1}{T_0}\pi_0(1 - \pi_0)\right] + \operatorname{var}(\pi_0) \doteq \frac{\pi_0(1 - \pi_0)}{p_0^{\star} N_0}.$$

Thus

$$\operatorname{var}(\hat{\pi}_0) \doteq \frac{\pi_0(1 - \pi_0)}{f_0 p_0^{\star} N_0}$$

leading to equation (1).

ACKNOWLEDGEMENTS

We thank Jeffrey L. Carson, M. D. and Amy Duff, M. H. S., for permission to use the liver disease example data.

REFERENCES

1. Carson, J. L. and Strom, B. L. 'Medicaid databases', *in* Strom, B. L. (ed.), *Pharmacoepidemiology*, 2nd edn, Wiley, Chichester, 1994, Chapter 15, pp. 199–216.
2. Strom, B. L., Tamragouri, R. N., Morse, M. L., Lazar, E. L., West, S. L., and Stolley, P. D. 'Oral contraceptives and other risk factors for gallbladder disease', *Clinical Pharmacology and Therapeutics*, **39**, 335–341 (1986).
3. Piper, J. M., Ray, W. A., Daugherty, J. R. and Griffin, M. R. 'Corticosteroid use and peptic ulcer disease: Role of nonsteroidal anti-inflammatory drugs', *Annals of Internal Medicine*, **114**, 735–740 (1991).
4. Brenner, H. and Gefeller, O. 'Use of the positive predictive value to correct for disease misclassification in epidemiologic studies', *American Journal of Epidemiology*, **138**, 1007–1015 (1993).

5. Carson, J. L., Strom, B. L., Duff, A., Gupta, A. and Das, K. 'Safety of nonsteroidal anti-inflammatory drugs with respect to acute liver disease', *Archives of Internal Medicine*, **153**, 1331–1336 (1993).
6. Carson, J. L., Duff, A. and Strom, B. L. 'Drug-induced acute liver disease', *Pharmacoepidemiology and Drug Safety*, **2**, S19–S23 (1993).
7. Ury, H., 'Efficiency of case-control studies with multiple controls per case: continuous or dichotomous data', *Biometrics*, **34**, 643–649 (1975).
8. West, S. L. and Strom, B. L. 'Validity of pharmacoepidemiology drug and diagnosis data', in Strom, B. L. (ed.), *Pharmacoepidemiology*, 2nd edn, Wiley, Chichester, 1994, pp. 549–580.
9. Brenner, H., Savitz, D. A. and Gefeller, O. 'The effects of joint misclassification of exposure and disease on epidemiologic measures of association', *Journal of Clinical Epidemiology*, **46**, 1195–1202 (1993).